IMPLEMENTATION OF HIDDEN MARKOV MODEL (HMM) FOR PARTS OF SPEECH TAGGING IN TELUGU LANGUAGE

Dr. V. Suresh

Assistant Professor in IT, Anil Neerukonda Institute of Technology and Sciences (ANITS), Sangivalasa, Visakhapatnam, Andhra Pradesh, India.

ABSTRACT

All NLP applications have fundamental task of POS(Parts of Speech) Tagging. Like Grammar Checking, Speech processing, Machine translation etc. that assign the correct tag to the word for a number of available tags. The accuracy of a tagger is the biggest challenge today. A lot of taggers have been proposed by different Researchers for the different languages (Telugu, Tamil, Kannada, Punjabi, Hindi, Bengali etc.) using different techniques like HMM (Hidden Markov Model), SVM (Support Vector Machine), ME (Maximum Entropy) etc. A Telugu POS tagger based on HMM model is one of them. This tagger uses Hidden Markov Model., a statistical technique to accurately tag the words in Telugu language using 630 tags developed by Rama Sree, R.J, Kusuma Kumari,2007.This large tag set (630 tags)results in data sparseness problem. Finally the result has been manually evaluated from a linguistic person. To cope up with this problem, in this research paper an experiment with reduced POS Tag set (36 tags) proposed by Technical Development of Indian Languages (TDIL) has been used to improve the tagging accuracy of HMM based POS Tagger.

KEYWORDS: Telugu, Parts-of-speech tagger, Corpus, TDIL proposed Telugu tag set, HMM technique.

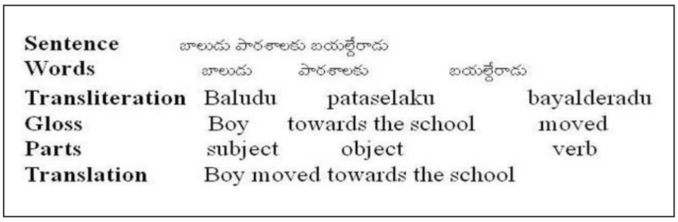
INTRODUCTION:

Telugu language belongs to Indo-Aryan family of languages (Dravidian languages). Other members that belong to this family are Kannada, Tamil, Malayalam, Hindi, Bengali, Gujarati, and Marathi etc. Telugu is spoken in India, Canada, USA, UK, and other countries with Telugu immigrants. Telugu language is the 8th most widely spoken language in the world, 4th most spoken language in Canada (The Times of India, 14th February, 2008) and the 9th in India with more than 45 million speakers. It is the official language of Telugu states (Andhra Pradesh and Telangana).

The first treatise on Telugu grammar, the "Andhra Shabda Chintamani" was written in Sanskrit by Nannaya who was considered as the first poet and translator of Telugu in the 11th century A.D. There was no grammatical work in Telugu prior to Nannayya's "Andhra shabda chintamani". This grammar followed the patterns which existed in grammatical treatises like Astadhvavi and Valmiki vyakaranam but unlike Paninni. Nannayya divided his work into five chapters, covering Samjna, Sandhi, Ajanta, Halanta and Kriya. After Nannayya, Atharvana and Ahobala composed Sutras, Vartika and, Bhashyam. Like Nannayya, they had previously written their works in Sanskrit.

In the 19th century, Chinnaya Suri wrote a simplified work on Telugu grammar called Balavyakaranam borrowing concepts and ideas from Nannayya's Andhra Shabda Chintamani, and wrote his literary work in Telugu. Every Telugu grammatical rule is derived from Paninian, Katvayana, and Patanjali concepts. However high percentage of Paninian aspects and techniques borrowed in Telugu.

According to Nannayya language without 'Niyama" or the language which doesn't adhere to Vyākaranam is called Gramya or Apabramsa and hence it is unfit for literary usage. All the literary texts in Telugu follows Vyākaranam. Compared to languages like English, Telugu is a morphologically rich language and has relatively free word order. It follows a Subject-Object-Verb (S-O-V) pattern. It is:



Sentence	బాలుడు పాఠశాలకు బయల్దేరాడు		
Words	బాలుడు	పాఠశాలకు	బయల్దేరాడు
Transliteration	Baludu	pataselaku	bayalderadu
Gloss	Boy	towards the school	moved
Parts	subject	object	verb
Translation	Boy moved towards the school		

This sentence can also be interpreted as 'Boy moved towards the school' depending on the context. But it does not affect the SOV order.

In most of the natural language processing applications like grammar checking, sentence identification, phrase chunking etc. the computer required only grammatical information of the input text. This grammatical information is given in the form tags called part of speech tags The process of assigning the correct tag to the word from a number of available tags is called POS Tagging. Here the grammatical information of the word is called Tag .It is well known that a computer will understand and process the language if the meaning of each and every word of that language is known. The parts of speech are different word classes in which a word lies like noun, adjective, verb etc. A word can occur in more than one word class in different context. Same word can act as a noun in one sentence and the same can act as a verb in other sentence.

So in order to assign the exact grammatical information to a word one must know the context in which that word has occurred. For a computer system it is very difficult to understand the context of the sentence. Therefore different techniques are used to assign the part of speech tag to a word.

RELATED WORK:

Different Techniques used for POS Tagging of Indian Languages:

There are basically three techniques used for part of speech tagging. 1) Rule based method 2) Statistical based method and 3) Neural network based method. Besides these three, a hybrid method is also used. This hybrid method is the combination of two or three of above mention techniques. In rule based technique different hand written rules are used for disambiguation of tags. These rules are developed manually. Therefore thorough knowledge of language is required to develop the rules. This rule based technique has been used by Sreeganesh (2006) for Telugu language; another rule based POS tagger was developed for Punjabi language by Mandeep Singh Gill, Gurpreet Singh Lehal (2008). Statistical method is another technique commonly used for part of speech tagging. Most commonly used statistical methods are support vector machine (SVM) used byEkbal and S. Bandyopadhyay(2008) for Bengali language; V.Dhanalakshmi et al. (2008)for Tamil language, M Anandkumar, Vijaya M.S, Loganathan R, Soman K.P, Rjendran S (2008) ;SindhiyaBinulal et al. for POS tagging of Telugu language. Antony P.J et al. for Malyalam language. Hidden markov model based technique used by Manish Shrivastava&Pushpak Bhattacharyya for POS tagger for Hindi language; Manju K et al. for Malayalam language; NavanathSaharia et.al for Assamese; Sanjeevkumar Sharma et al. (2011) for Punjabi Language; Ekbal, S. Mondal et al. for Bengali language. Maximum entropy based technique was used by AniketDalal et al. for Hindi language; Ekbal et al. (2008) for Bengali language.

Conditional Random Field based technique has been used by Ravindran et al. and Himanshu et al. for POS tagging and chunking of Hindi language; other Indian languages on which this CRF technique has been applied are Bengali and Manipuri. Neural network based technique has been used by Ankur Parikh for Hindi Language. In hybrid based approach used a combination of rule based and HMM based technique has been used by Arulmozhi P et al. for development of Tamil POS tagger; Chirag Patel and KarthikGali used a combination of rule based method and CRF for Gujarati Pos tagger

Existing POS Tagger of Telugu Language:

In Indo-Aryan family, Telugu language is a most popular language. It is also known as Indian language. Other members of this family are Hindi, Bengali, Gujarati, and Marathi, ,Punjabi etc. Two POS tagging system with two different techniques have been developed for Telugu language .by RamaSree, R. J., and P. Kusuma Kumari in .2007. The first system was developed as a sub part of grammar checker project. These rules were implemented by using regular expression.

The main reason for using this rule based technique was that the rules can be edited i.e. new rules can be added or deleted.

A tag set of more than 630 tags was also developed. Second POS tagging system has been developed by using statistical method. Hidden Markov model was used to disambiguate the tags. Viterby algorithm was used for implementation of Hidden Markov model. The tag set used in the second system was same as was proposed by Mandeep singh et al. They also tried a hybrid approach that is combination of rule based system and statistical approach in which the output of rule based system was fed to the statistical based system. This gives further improvement on the accuracy of the POS tagger.

Tag set:
The grammatical information of the language represents a set of all the tags used in a tag set. The number of tags used for a language depends upon the length of the tag which further depends upon the amount of information that we want to represent using a tag. e.g. if just basic word class is to be represented with each word then the length of the tag will be 2, 3 or 4. One extra character will be required for extra grammatical information that is to be represented with tag. E.g. to represent only word class we can use NN tag for noun. But if we want to represent gender information also then an extra character will be added to this tag. This extra character may be M for masculine gender, F for feminine gender and B for both types. Therefore proposed tag for a masculine noun will be NNM, for feminine it will be NNF and for both categories it may be NNB. This extra information not only increases the length of the tag but also increase the no of tags. As in above case if the information of only word class is to be represented then only one tag was sufficient and as the information increases the number of tags also increase and becomes 3 in above case. From above discussion it is concluded that the information has a direct effect on the number of tags.

Telugu tag set:
Existing Telugu POS Tag set:
Two POS tagger has been developed for Telugu language are discussed in above section. In both of these POS taggers same tag set has been used. This tag set was developed by keeping in mind that this POS tagger has to be used for grammar checking software of Telugu language. This tag set was fine grained and more than 630 tags were used.

New proposed Tag set by TDIL:
A number of POS tag sets have been developed by different organizations are depending on some general principle of tag set design strategy. For POS annotation of texts in Telugu, we have used tag set proposed by TDIL (Technical Development of Indian Languages). There were 36 tags proposed by TDIL for Telugu language.

INTRODUCTION TO HMM (Hidden Markov Model):
Hidden Markov Model is a statistical model used to solve classification type of problems. It was proposed by L. E. Baum. This model is used to assigns the joint probability to paired observation and label sequence. In order to maximize the joint likelihood of training sets parameters are trained. In NLP applications this type of training is done by using accurate annotated corpus. The main advantage of this model is that it is easy to understand and implemented. The accuracy of this model is directly proportional to size of training data.

Basic Definitions and Notation:
According to (Rabiner, 1989), the HMM can be defines by using the following five elements:

1. N, it is the number of distinct states in the model. For part-of-speech tagging, N is the total number of tags that can be used by the system. In existing system these are more than 360 tags and in our propose system these are 36 tags only. Each possible tag for the system corresponds to one state of the HMM.

2. M, the number of distinct output symbols in the alphabet of the HMM. For part-of-speech tagging, M is the number of words in the lexicon of the system. As the exact number of words of a language can't be counted so the distinct words present in the training corpus is taken as M.

3. A = {a_{ij}}, the state transition probability distribution. Where $a_{ij}$is the probability that the system will move from state i to state j in one transition. For part-of-speech tagging, these states are represented by the tags, so $a_{ij}$is the probability that the model will move from tag t_i to t_j. This probability can be estimated using training corpus.

4. $B = \{b_j(k)),$ the emission probability. The probability $b_j(k)$is the probability that the k-th output symbol will be emitted when the model is in state j. For part-of-speech tagging, this is the probability that the word $W_k$will be assign tag t_j(i.e.,$P(W_k|t_j))$. This probability can be again estimated from a training corpus.

5. $\prod = \{\prod_i\}$, the initial state distribution. It is the probability that the model will start in state i. For part-of-speech tagging, this is the probability that the sentence will begin with tag ti. When using an HMM to perform part-of speech tagging, the goal is to determine the most likely sequence of tags (states) that generates the words in the sentence (sequence of output symbols). In other words, given a sentence V, calculate the sequence U of tags that maximizes P (V/U). The Viterbi algorithm is a common method for calculating the most likely tag sequence when using an HMM. The proposed model is a type of first order HMM, also referred to as bigram POS tagging. For POS-tagging problem presented Hidden Markov Model is composed of two probabilities: lexical (emission) probability and contextual (transition) probability (Samuelsson, 1996).

$$(t_1,\ldots,t_n)^* = \underset{t_1\ldots t_n}{\operatorname{argmax}}\ P(t_1,\ldots,t_n)|(w_0,\ldots,w_n)$$

Using Baye's law above equation can be rewritten as:

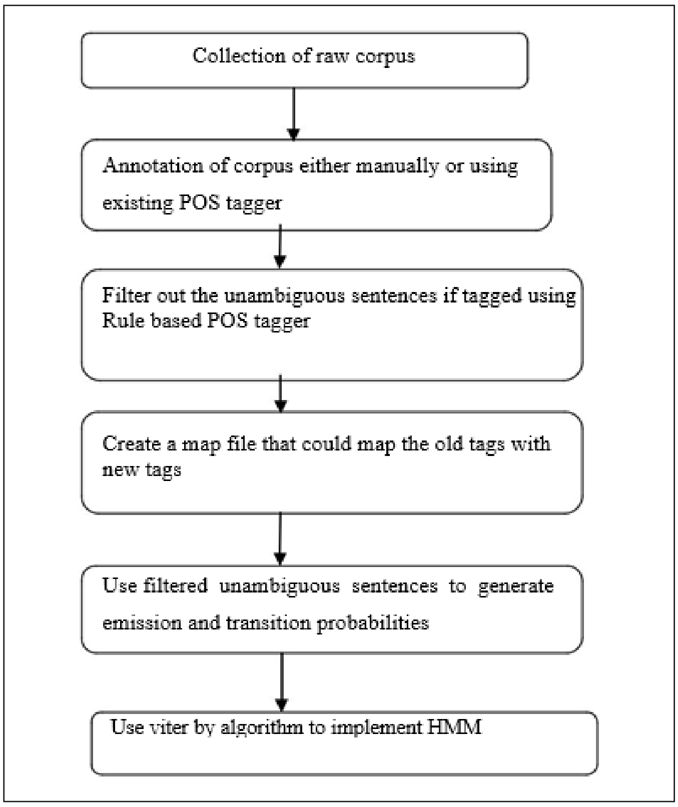
$$P(t_1,\ldots\ldots\ldots t_n|W_1,\ldots\ldots\ldots,W_n)=P(t_1,\ldots\ldots\ldots\ldots\ldots t_n)\ X\ \frac{P(W_1,\ldots\ldots,W_n|t_1,\ldots\ldots,t_n)}{P(W_1,\ldots\ldots.,W_n)}$$

$$(t_1,\ldots\ldots\ldots.,t_n)=\ \arg\ \max P(t_1,\ldots.,t_m)\ X\ P\ (\ _1,\ldots\ldots\ldots..W_n|t_1,\ldots\ldots..t_n)$$

$$(t_1,\ldots\ldots\ldots.,t_n)= \arg\ \max P(t_1,\ldots\ldots\ldots,t_{\ m})|\ (W_0,\ldots\ldots\ldots,W_n)$$

$$= argmax\Pi(P(ti|ti-1)(TRANSITION\ PROBABILITY) * P(Wi|Ti)(EMMISION\ PROBABLITY))$$

METHODOLOGY:
Following flow diagram shows the steps used to implement new tag set in HMM.


Collection of raw corpus
Annotation of corpus either manually or using existing POS tagger
Filter out the unambiguous sentences if tagged using Rule based POS tagger
Create a map file that could map the old tags with new tags
Use filtered unambiguous sentences to generate emission and transition probabilities
Use viter by algorithm to implement HMM

Step 1:- collection of raw corpus.
We collected a large accurate corpus of nearly 200 pages containing nearly 8000 sentences and approximately 42,000 words. This corpus was collected from internet. Following web sites were used for the collection of corpus:

- http://telugutribuneonline.com
- www.teluguinfoline.com

Step 2:- Annotation of corpus.
For the annotation we used an existing lexicon based morphological analyzer. This morphological analyzer contains more than one million Telugu words with their part of speech tag.

Step 3:- Mapping with new tags.
Since the existing morph used in step 2 contains tags that have been developed using a tag set of 630 tags and in new system we have to use the tags proposed by TDIL, so mapping of tags was done to reduce the

630 tags to 36 tags. The tags of annotated corpus developed in above step were converted in to standard common tags set for Indian Languages and as per IIIT tag set guidelines. This mapping was partially done manually and partially by computer by using a map file. The map file was manually developed with the help of linguistic.

Mapped File

S. No		Old tag starts with	New tag
1		NN	N_NN/N_NST/N_NNP
2		PNP	PR_PRP
3		PNR	PR_PRF
4		PND	PR_PRC
5		PNI	PR_PRI
6		PNE	PR_PRL
7		PNN	PR_PRQ
8		AJI	JJ
9		CD	QT_C
10		OD	QT_O
11		PPU	PSP
12		AVI	RB
13		PPI	PSP
14		AVU	RB
15		VBP	V
16		CJ	CC
17		PTU	RP
18		PTV	RP
19		AJU	JJ
20		VBO	V
21		VBMA	V_VM
22		BVAX	V_VAUX
23	2286	Unknown	RD_UNK
24	Comma , Dot,	Question Sentence, Sentence, Exclamation	
	Colon, Semicolon, Opening Single Quote, Closing Single Quote,		
	Opening Double Quote, Closing Double Quote	RD_PUNC	
25	Opening Bracket Opening Brace, Closing Bracket, Closing Brace		
	Opening Parenthesis, Closing Parenthesis, Less Than, Greater Than	RD_SYM	
26		VBMAXPSXXTNE	V_VM_VNG
27		AJU	RP__INTF
28		PTUN	RP__NEG
29		AJIMSD	QT__QTF
30.		No specific match but generally unknown words	RD_RDF
31		No specific tag but words with hyphen in between them	RD_ECH
32		VBMAXXXXXINIAN	V__VM__VINF
33		VBMAXXXXXINDIAN	V__VM__VNF

Step 4:- Development of emission and transition probability file. Now in order to find out transition and emission probabilities we developed an application in visual studio (c#.net). The probability files were kept in txt format. These file were used for implementing HMM using viterby algorithm

Sample transition file

Tag1/Tag2 pair	Probability
N_NN/V	0.190476
V/V_VM	0.005376
V_VM/RP	0.040698
RP/PR_PRP	0.016484
PR_PRP/N_NN	0.01066
N_NN/PSP	0.470769
PSP/N_NN	0.028161
N_NN/V_VM	0.135289

Step 5:- Viterby algorithm was used to implement HMM.

Experimental Evaluation:

The accuracy of Natural Language product is generally measured in terms precision and recall. Precision is the percentage of correctly disambiguate tags. And recall is If A is the number of correctly disambiguate tags and B is the number of tags that were not disambiguate by our system then

$$\text{Recall} = A / (A+B)$$

Similarly if A is the no of correctly disambiguated tags and C is the number of incorrectly disambiguated tags then Precision= A / (C+A)

For evaluation of the proposed POS Tagger, a corpus having texts from different online resources i.e. Telugu websites were used. The outcome was manually evaluated through a linguistic expert to mark the correct and incorrect disambiguate tags. The result obtained has been given in Table 1. The precision and recall values are given in table 2.

Table 1: Experimental Result

Corpus	Total number of words	No of unknown words (not tagged by the system)	No of known words	Existing HMM based system	Proposed system
				No of correctly disambiguated tags	No of correctly disambiguated tags
Essay	6894	442	6420	6132	6409
News	5387	384	4853	4803	4783
Short stories	7639	132	8324	8146	8231
Novel	2705	322	3754	3765	3268
Book Chapter	3876	74	3765	3754	3174

Table 2: Precision and Recall of HMM based and Proposed system

Corpus type	Existing HMM based system					Proposed System				
	A	B	C	Precision	Recall	A	B	C	Precision	Recall
Essay	6548	846	0	100%	84.5%	6528	0	71	98.8%	100%
News	4723	537	0	100%	85%	4698	0	70	98.4%	100%
Short stories	9573	2286	0	100%	85.5%	9297	0	47	98.2%	100%
Novel	3287	345	0	100%	86%	3387	0	40	99.3%	100%
Book Chapter	3864	378	0	100%	86.5%	3184	0	34	98.7%	100%

CONCLUSION:

Hidden Morkov Model (HMM) for Parts of Speech Tagging is of implementation of reducing the tag set has been done by efforts to improve the accuracy of HMM based Telugu POS Tagger. The tag set has been reduced from more than 630 tags to 36 tags. We observed a significant improvement in the accuracy of tagging. Our proposed tagger shows an accuracy of 95-98% whereas the existing HMM based POS Tagger was reported to give an accuracy of 84-88%. This significant improvement is due to reduction in the tag set from more than 630 tags to 36 tags. The main problem with large tag set results in data sparseness. The reduction in the tags results in reduction in data sparseness and hence improves the accuracy and enhancement.

Future scope:

The usage of a hybrid model i.e. combination of more than one statistical model can be further extended. The same approach can also be implemented for different Indian languages like Hindi, Bengali, and Punjabi etc. Applying different approaches to find the POS tag of unknown words are yet to done for improvement.

REFERENCES:

I. Ahmed, Raju S.B, Chandrasekhar Pammi V. S.,Prasad M.K (2002), "Application of multilayerperceptron network for tagging parts-of- speech", Proceedings of the Language EngineeringConference, IEEE.

II. RamaSree, R. J., and P. Kusuma Kumari.2007. Combining POS taggers for improved accuracy to create telugu annotated texts for information retrieval. Dept. of Telugu Studies, Tirupathi, India (2007).

III. Krishnapriya, V., Sreesha, P., Harithalakshmi, T. R., Archana, T. C., & Vettath, J. N. 2014. Design of a POS tagger using conditional random fields for Malayalam. IEEE 2014 First International Conference on Computational Systems and Communications (ICCSC), pp. 370-373

IV. Sharma S.K, Lehal G.S (2011) "Using HMM to Improve accuracy of Punjabi POS tagger" 2011 IEEE International Conference on computer science and Automation Engineering. Shanghai (China)

V. Kumar, D., & Josan, G. S. 2010. Part of speech taggers for morphologically rich Indian languages: a survey. International Journal of Computer Applications (0975-8887) Volume, 1-9.

VI. Reddy, S., & Sharoff, S. 2011. Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. Cross Lingual Information Access, p. 11.

VII. AniketDalal, Kumar Nagaraj, Sawant Uma,ShelkeSandeep (2006), "Hindi Part-of-SpeechTagging and Chunking: A Maximum Entropy Approach" Proceedings of the NLPAI MLcontestworkshop, National Workshop on Artificial Intelligence.

VIII. Ankur Parikh (2009), "Part-Of-Speech Tagging usingNeural network", Proceedings of ICON-2009: 7th International Conference on Natural Language Processing.

IX. Manju, K., Soumya, S., & Idicula, S. M. 2009. Development of a POS tagger for Malayalam-an experience. IEEE International Conference on Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. pp. 709-713.

X. Antony P.J, Mohan S. P., Soman K.P (2010), "SVM Based Part of Speech Tagger for Malayalam", Proceedings of 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, IEEE.

XI. AnupamBasu, Ray, RanjanPradipta, Harish V.and SarkarSudeshna(2003), "Part of speech tagging and local word grouping techniques fornatural language parsing in Hindi", Proceedings of the International Conference on Natural Language Processing (ICON 2003).

XII. Arulmozhi.P, L Sobha (2006) "A Hybrid POS Tagger for a Relatively Free Word Order Language",Proceedings of MSPIL-2006, Indian Institute of Technology, Bombay.

XIII. Avinesh PVS and GaliKarthik (2007), "Part-of-speech tagging and chunking using conditional random fields and transformation based learning", Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL), pp. 21–24.

XIV. Ekbal, S. Mondal and S. Bandyopadhyay (2007). POS Tagging using HMM and Rule-based Chunking. In Proceedings of the Workshop on Shallow Parsing in South Asian Languages, International Joint Conference on Artificial Intelligence (IJCAI 2007), 6-12 January 2007, Hyderabad, India, PP. 25-28.

XV. Ekbal, R. Haque and S. Bandyopadhyay (2007), "Bengali Part of Speech Tagging using Conditional Random Field", Proceedings of the 7th International Symposium on Natural Language Processing (SNLP-07), Thailand, pp.131-136.

XVI. Ekbal and S. Bandyopadhyay (2008), "Part of Speech Tagging in Bengali using Support Vector Machine",Proceedings of the International Conference on Information Technology (ICIT 2008), pp.106-111, IEEE.

XVII. Ekbal, M. Hasanuzzaman and S. Bandyopadhyay (2009), "Voted Approach for Part of speech Tagging in Bengali", Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC-09), December 3-5, Hong Kong, pp. 120-129.

XVIII. Ganesan M (2007), "Morph and POS Tagger for Tamil" (Software) Annamalai University, Annamalai Nagar.

XIX. G.SindhiyaBinulal, Goud P. A, K.P.Soman(2009), "A SVM based approach to Telugu Parts of Speech Tagging using SVMTool", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

XX. HimanshuAgrawal, Mani Anirudh (2006), "Part Of Speech Tagging and Chunking Using Conditional Random Fields" Proceedings of the NLPAI MLcontest workshop, National Workshop on Artificial Intelligence.

XXI. Mandeep Singh Gill, Lehal G.S. (2008) "Grammer Checking System for Punjabi" Coling 2008:companion volume Posters and Demonstrations pages 149–152 Manchester.

XXII. Manish Shrivastava, Bhattacharyya Pushpak (2008), "Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge", Proceedings of ICON-2008: 6th International Conference on Natural Language Processing.

XXIII. Manjuk,SSoumya, Idicula S.M. (2009), "Development of A Pos Tagger for Malayalam-An Experience", Proceedings of 2009 International Conference on Advances in Recent Technologies in Communication and Computing, IEEE .

XXIV. NavanathSaharia, Das Dhrubajyoti, Sharma Utpal, KalitaJugal (2009), "Part of Speech Tagger for Assamese Text", Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec, Singapore, pp. 33–36

XXV. Ratnaparkhi, A. 1996. A maximum entropy model for part-of-speech tagging. In Proceedings of the conference on empirical methods in natural language processing Vol. 1, pp. 133-142.